NONPARAMETRIC BAYES I

JOE WENJIE JIANG, MIKE WAWRZONIAK

1. INTRODUCTION

In finite mixture models, we know a priori the number K of clusters existing in the data. Each data point is generated by one of K distributions, each of which is characterized by some parameters. For example, we can cluster the data using K-means or Gaussian mixture models. These approaches are widely used in machine learning and statistics, and are applied in areas such as image processing, information retrieval and gene expression analysis. A fundamental question that naturally arises is how to do model selection, i.e., how can we choose right number of clusters.

Bayesian nonparametrics provides a form of model selection and a flexible model. Nonparametric does not mean that there are no parameters. It simply means that the number of parameters in the model can grow as we get more data. This allows our model the flexibility to be as complex as our data needs. The Dirichlet process introduced by Ferguson 1973 and Antoniak 1974 provides such machinery to let K grow with the data. The data analysis usually relies on MCMC to obtain an estimate of parameters in the model.

2. CHINESE RESTAURANT PROCESS

There are many ways of looking at the Dirichlet process. We start by examining the Chinese restaurant process (CRP) representation and see its connection with the Dirichlet process. The Chinese restaurant process is a distribution of partitions of integers, which is introduced by Pitman and Dubins. Suppose N customers arrive at a restaurant with infinite capacity sequentially. Denote n_i as the number of customers already sitting in table i. Each incoming customer chooses a table at random, with probability that

$$\text{(1)} \qquad \begin{cases} p(\text{table } i|\text{previous customers}) & \propto \quad n_i \\ p(\text{next empty table}|\text{previous customers}) & \propto \quad \alpha \end{cases}$$

where α is a parameter that is invariable during the process. Note that a new customer can either sit at an existing table, or can start a new table. Figure 1 illustrates an example of the Chinese restaurant process, in which each customer is numbered by his arrival sequence.

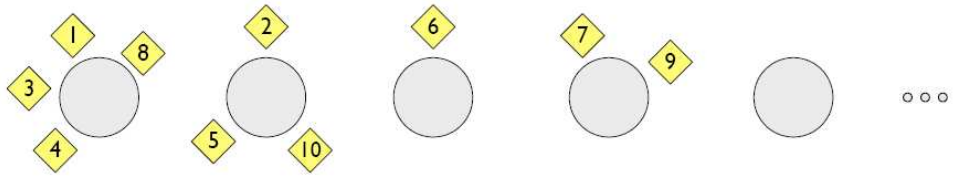

FIGURE 1. An Example of the Chinese Restaurant Process

The Chinese restaurant process is a representation of partition of integers $\{1, 2, \ldots, N\}$. For example, in Figure 1, 10 numbers are partitioned into four groups $(1, 3, 4, 8), (2, 5, 10), (6)$ and $(7, 9)$. The partition is free of orders, i.e., we can renumber the tables to produce the same partition. The expected number of tables taken by N customers is

$$E[\text{\# of tables}] = \alpha \log N \tag{2}$$

In CRP mixture model, each data point is generated as follows. Each table corresponds to a cluster, which is associated with a parameter drawn from a prior $p(\eta^*|\lambda)$. For each customer, we first choose a table $Z \sim \text{CRP}(\alpha)$. Then we draw a value from $p(x|\eta^*_z)$, e.g., a prior over Gaussian locations. Such a process is illustrated in Figure 2. The generative process is similar to mixture models, but with unbounded number of mixture components. Note that an outlier in the data can be thought of as a new cluster. So the the emphasis of CRP is not on the actual number of clusters.

Given data $\{x_1, \ldots, x_N\}$, the posterior is a distribution on

- number of clusters (number of occupied tables)
- which data are assigned to each cluster
- parameter η^*_z of each cluster

Generally, the number of clusters is random and unknown, and new data can be assigned to a new cluster. Our goal is to estimate this posterior distribution. The CRP mixture model is well studied in statistics, e.g., Escobar and West 1994, Neal 2001, Gelfand and Kottas 2002. It is widely used in many application such as spatial statistics, computer vision and censored models.

There are several useful extensions of CRP developed in machine learning. One famous example is nested CRP, which is shown in Figure 3. In nested CRP, there can be infinite levels of CRPs recursively defined. For example, at the top level, each table is associated with a CRP parametered by η^*. Customers on a table can be further clustered into smaller tables, and this process can continue infinitely. With nested CRP, what you finally construct is a random tree. This is exactly the powerful expressiveness of

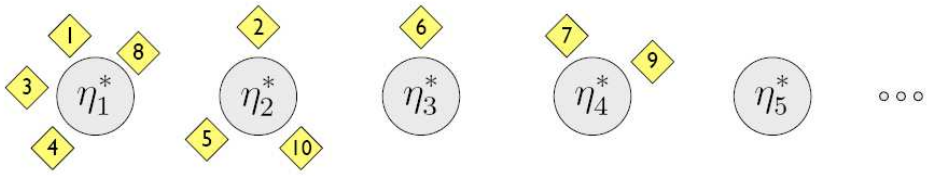

FIGURE 2. Generative process of CRP mixtures

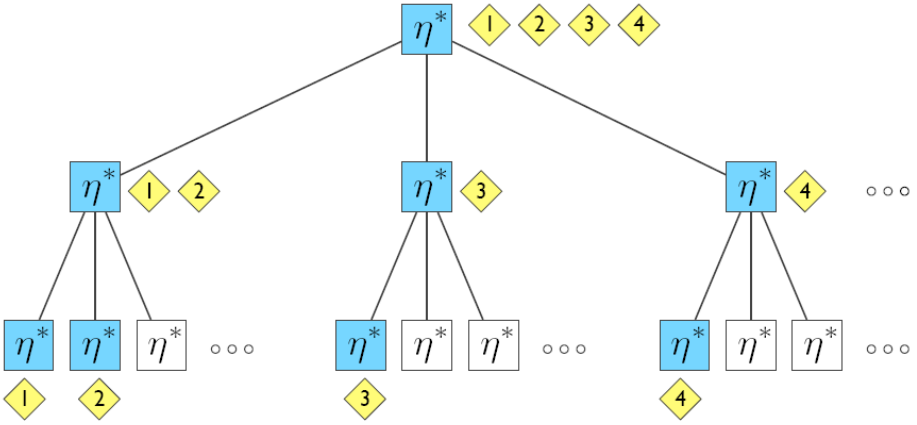

FIGURE 3. Nested CRP

nested CRP. CRP is a simple distribution over partition, while nested CRP is a distribution over tree, with which one can search over complicated combinatorial structures. One example of such application is to classify the key words of computer science journal articles into hierarchical structures, such as a tree shown in Figure 4. Each node on the same depth represents the key words of a branch area, and this branch can be further classified into sub areas.

3. THE STICK-BREAKING CONSTRUCTION

An alternative perspective on the DP is one based on the stick-breaking construction due to Sethuraman [1]. The representation for a $\mathrm{G} \sim \mathrm{DP}(\alpha_0, G_0)$ is constructed from an infinite sequence of iid random variables π'_k and ϕ'_k where $k = 1 \ldots \infty$,

$$\begin{aligned} \pi'_k \mid \alpha_0 &\sim \mathrm{beta}(1, \alpha_0) \\ \phi_k \mid \mathrm{G}_0 &\sim \mathrm{G}_0. \end{aligned} \tag{3}$$

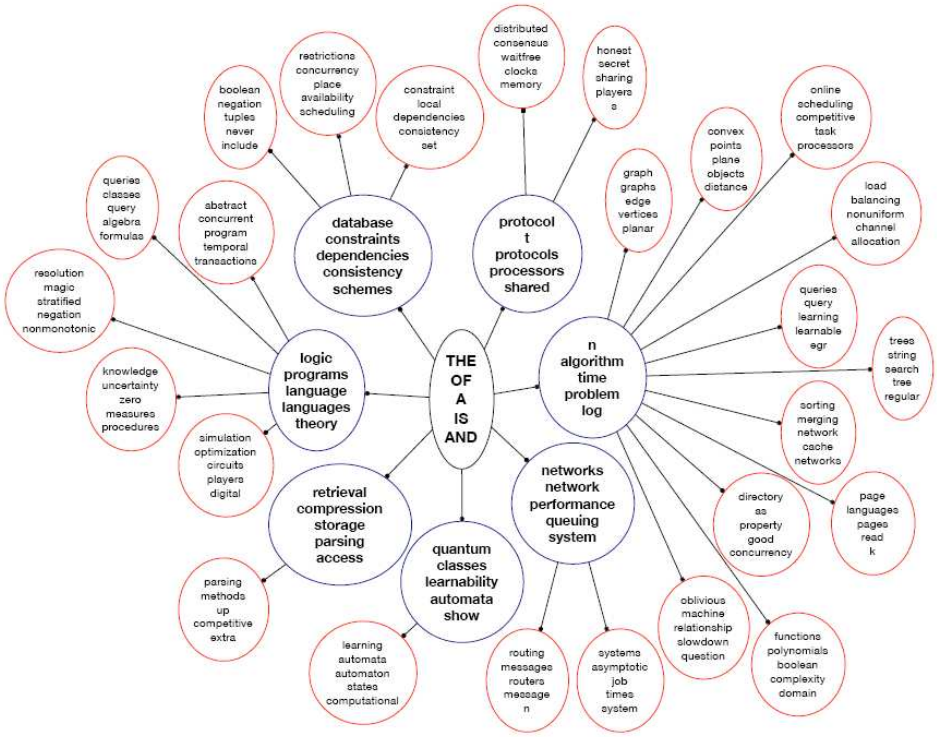

FIGURE 4. Analyzing key words of computer science journal articles using nested CRP

The actual measure G is then derived from the random variables as follows

$$\begin{aligned} \pi_k &= \prod_{l=1}^{k-1}(1-\pi'_l) \\ G &= \sum_{k=1}^{\infty}\pi_k\delta_{\phi_k}, \end{aligned} \tag{4}$$

where δ_{ϕ_k} is a probability measure concentrated at ϕ_k and $\sum_{k=1}^{\infty}\pi_k = 1$. It can be seen from this construction that $\boldsymbol{\pi}$ is a random probability measure on positive integers, and measures drawn from DP are discrete with probability 1. Figure 5 is a prictorial representation of a unit stict constructed according to the stick-breaking procedure. Probabilty measure $\boldsymbol{\pi}$ defined by Equation 4 is also refered to as $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0)$.

The sitck-breaking construction can be useful in definig more complex models, such as the Hierarchical Didirchlet Process (HDP) [2]. The HDP

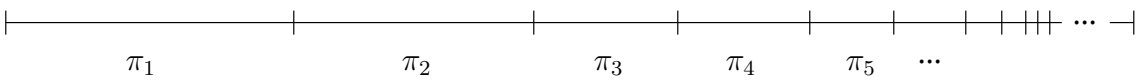

FIGURE 5. A unit length stick representation, $\sum_{k=1}^{\infty} \pi_k = 1$.

construction can also be represented in the stick-breaking interpretations as

$$\begin{aligned} \beta &\sim GEM(\gamma) \\ \theta_i^{**} &\sim H \\ G_0 &= \sum_{i=1}^{\infty} \beta_i \delta(\theta_i^{**}) \\ \pi_j &\sim GEM(\alpha) \\ \theta_{ji}^{*} &\sim G_0 \\ G_j &= \sum_{i=1}^{\infty} \pi_{ji} \delta(\theta_{ji}^{*}) \ . \end{aligned} \tag{5}$$

Because G_0 has support at the points θ_i^{**}, G_j has support at these points as well, and therefore can also be written as

$$G_j = \sum_{i=1}^{\infty} \omega_{ji} \delta(\theta_i^{**}) \ . \tag{6}$$

Since $G_j \sim \mathrm{DP}(\alpha, G_0)$, then for a measurable partition $(A_1, \ldots, A_r)$ of Θ

$$(G_j(A_1), \ldots, G_j(A_r)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_r)) \tag{7}$$

Therefore, if for $l = 1, \ldots, r$ let $I_l = \{i : \theta_i^{**} \in A_l\}$

$$\left(\sum_{i \in I_1} \omega_{ji}, \ldots, \sum_{i \in I_r} \omega_{ji} \right) \sim \mathrm{Dir}\left(\alpha \sum_{i \in I_1} \beta_i, \ldots, \alpha \sum_{i \in I_r} \beta_i \right) \tag{8}$$

Hence, $\omega_j \sim \mathrm{DP}(\alpha, \beta)$.

The random probability measure ω_j is also produced with the stick-breaking construction

$$\begin{aligned} \omega'_{ji} &\sim \mathrm{beta}\left(\alpha\beta_i, \alpha\left(1 - \sum_{l=1}^{i} \beta_l \right) \right) \\ \omega_{ji} &= \omega'_{ji} \prod_{l=1}^{i-1} (1 - \omega'_{jl}) \ , \end{aligned} \tag{9}$$

and also by

$$\omega_{ji} \sim \mathrm{beta}(\alpha\beta_i, \alpha(1 - \beta_i)) \ . \tag{10}$$

To arrive at Eq. 10 and Eq. 9, note that for partition $(1, \ldots, i-1, i, i+1, i+2, \ldots)$ by Eq. 8

$$\left(\sum_{l=1}^{i-1} \omega_{jl}, \omega_{ji}, \sum_{l=i+1}^{\infty} \omega_{jl}\right) \sim \mathrm{Dir}\left(\alpha \sum_{l=1}^{i-1} \beta_l, \alpha\beta_i, \alpha \sum_{l=i+1}^{\infty} \beta_l\right) \tag{11}$$

Eq. 10 follows by standard properties of Dirichlet distribution. Also by standard properties of Dirichlet distribution,

$$\left(\frac{\omega_{ji}}{1 - \sum_{l=1}^{i-1} \omega_{jl}}, \frac{\sum_{l=i+1}^{\infty} \omega_{jl}}{1 - \sum_{l=1}^{i-1} \omega_{jl}}\right) \sim \mathrm{Dir}\left(\alpha\beta_i, \alpha \sum_{l=i+1}^{\infty} \beta_l\right). \tag{12}$$

Then defining,

$$\omega'_{ji} = \frac{\omega_{ji}}{1 - \sum_{l=1}^{i-1} \omega_{jl}}, \tag{13}$$

and therefore,

$$\omega_{ji} = \omega'_{ji} \prod_{l=1}^{i-1} (1 - \omega'_{jl}). \tag{14}$$

Together with

$$1 - \sum_{l=1}^{i} \beta_l = \sum_{l=i+1}^{\infty} \beta_l \tag{15}$$

arrive at Eq. 9.

The concentration parameters γ, α and the baseline probability measure H are the hyperparameters of an HDP. For small values of γ, the mass is concentrated on a few atoms of H, as the value of γ increases, mass shifts away to be more spread out. This can be observed on the top row of figure 6 with draws for different values of γ. Similarly, for small values of α, mass is concentrated on few atoms of G_0, and as it increases the mass is less concentrated and more spread out. Since the distribution of G_0 is govern by the parameter γ, the parameter α can be interpreted as a refinement of the concentration on H set by γ. This can be seen in figure 6 with draws of ω for different values of α given a particular draw of β.

To formulate the HDP mixture model in the stick-breaking representation, let θ_{ji} be the factors corresponding to a single observation x_{ij}, and let :

$$\begin{aligned} \theta_{ji} &\sim G_j \\ x_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

where $F(\theta_{ji})$ denotes the distribution of the observation x_{ji}.

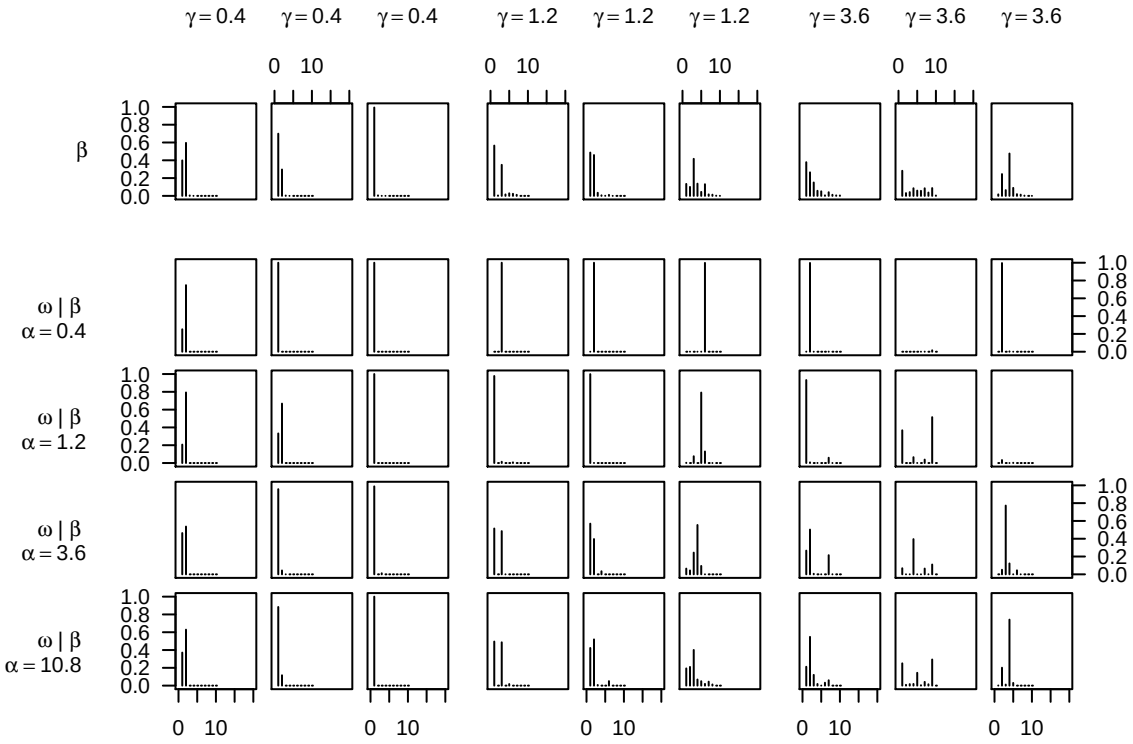

FIGURE 6. Random draws for β and γ. The top row shows 9 draws for β for γ of $0.4, 1.2$ and 3.6. Below each draw of β, 4 draws for ω are shown given the draw of β.

REFERENCES

[1] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.